Data Analysis Project

Introduction

We will be working with a dataset of auto claims filed by customers of an automobile insurance company located in the southwest and western regions of the United States.

Insurance companies depend on accurate pricing estimates to maintain profitability. Auto policies must be priced so that the insurance company makes a profit in the long run, given the costs of their customers' payouts for accident repairs, total loss car replacements, medical fees, and legal fees.

The executives at this insurance company have noticed declining profitability over the last several years and have hired you as a data science consultant to evaluate their claims data and make recommendations on pricing, customer behavior, and car insurance policy adjustments.

The objective of this project is to perform an exploratory data analysis on the claims_df dataset and produce an executive summary of your key insights and recommendations to the executive team at the insurance company.

Before you begin, take a moment to read through the following insurance company terms to familiarize yourself with the industry: Auto Insurance Terms (https://www.iii.org/article/auto-insurance-jargon-buster)

Auto Claims Data

The `claims_df` data frame is loaded below and consists of 6,249 auto claims submitted by customers of the insurance company. The rows in this data frame represent a single claim with all of the associated features that are displayed in the table below.

Note: If you have not installed the `tidyverse` package, please do so by going to the `Packages` tab in the lower right section of RStudio, select the `Install` button and type `tidyverse` into the prompt. If you cannot load the data, then try downloading the latest version of R (at least 4.0). The `readRDS()` function has different behavior in older versions of `R` and may cause loading issues.

```
installed.packages("tidyverse")
```

```
##      Package LibPath Version Priority Depends Imports LinkingTo Suggests
##      Enhances License License_is_FOSS License_restricts_use OS_type Archs
##      MD5sum NeedsCompilation Built
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6     ✔ purrr   0.3.4
## ✔ tibble  3.1.8     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
claims_df <- readRDS(url('https://gmubusinessanalytics.netlify.app/data/claims_df.rds'))
```

Raw Data

```
claims_df
```

customer_id <chr>	customer_state <fct>	highest_education <fct>	employment_status <fct>	gen… <fct>	inco… <dbl>	residenc <fct>
AA11235	Nevada	Bachelor	Medical Leave	Female	11167	Suburbar
AA16582	Washington	Bachelor	Medical Leave	Male	14072	Suburbar
AA34092	California	Associate	Employed	Male	33635	Suburbar
AA56476	Arizona	High School	Employed	Female	74454	Suburbar
AA69265	Nevada	Bachelor	Employed	Female	60817	Suburbar
AA71604	Arizona	Master	Employed	Female	87560	Suburbar
AA93585	California	Associate	Employed	Male	97024	Urban
AB21519	California	Associate	Employed	Female	93272	Urban
AB23825	California	Associate	Employed	Male	21509	Suburbar
AB26022	Oregon	High School	Retired	Male	26487	Suburbar

1-10 of 6,249 rows | 1-7 of 20 columns Previous **1** 2 3 4 5 6 … 625 Next

Exploratory Data Analysis

Executives at this company have hired you as a data science consultant to evaluate their claims data and make recommendations on pricing, customer behavior, and car insurance policy adjustments.

You must think of **at least 8 relevant questions** that will provide evidence for your recommendations.

The goal of your analysis should be discovering which variables drive the differences between customers with large lifetime values and customers who cost the company more than they provide in revenue through monthly premiums.

Some of the many questions you can explore include:

- Are there types of customers, based on their policy or demographics, that are highly profitable?
- Do certain policies have a lower number of claims, leading to large profits?
- Are there “problem customers” which have a large number of claims?

You must answer each question and provide supporting data summaries with either a summary data frame (using `dplyr` / `tidyr`) or a plot (using `ggplot`) or both.

In total, you must have a minimum of 5 plots and 4 summary data frames for the exploratory data analysis section. Among the plots you produce, you must have at least 4 different types (ex. box plot, bar chart, histogram, heat map, etc…)

Each question must be answered with supporting evidence from your tables and plots.

Question: Are there types of customers, based on their policy or demographics, that are highly profitable?

Answer: In comparison to the other two products, we can observe that the majority of clients choose personal insurance. However, we insurance companies earn more from corporate policy holders since the average customer life value is higher than the other two plans, as can be shown. Furthermore, the std of corporate policy is high, with a value of 2077.395.

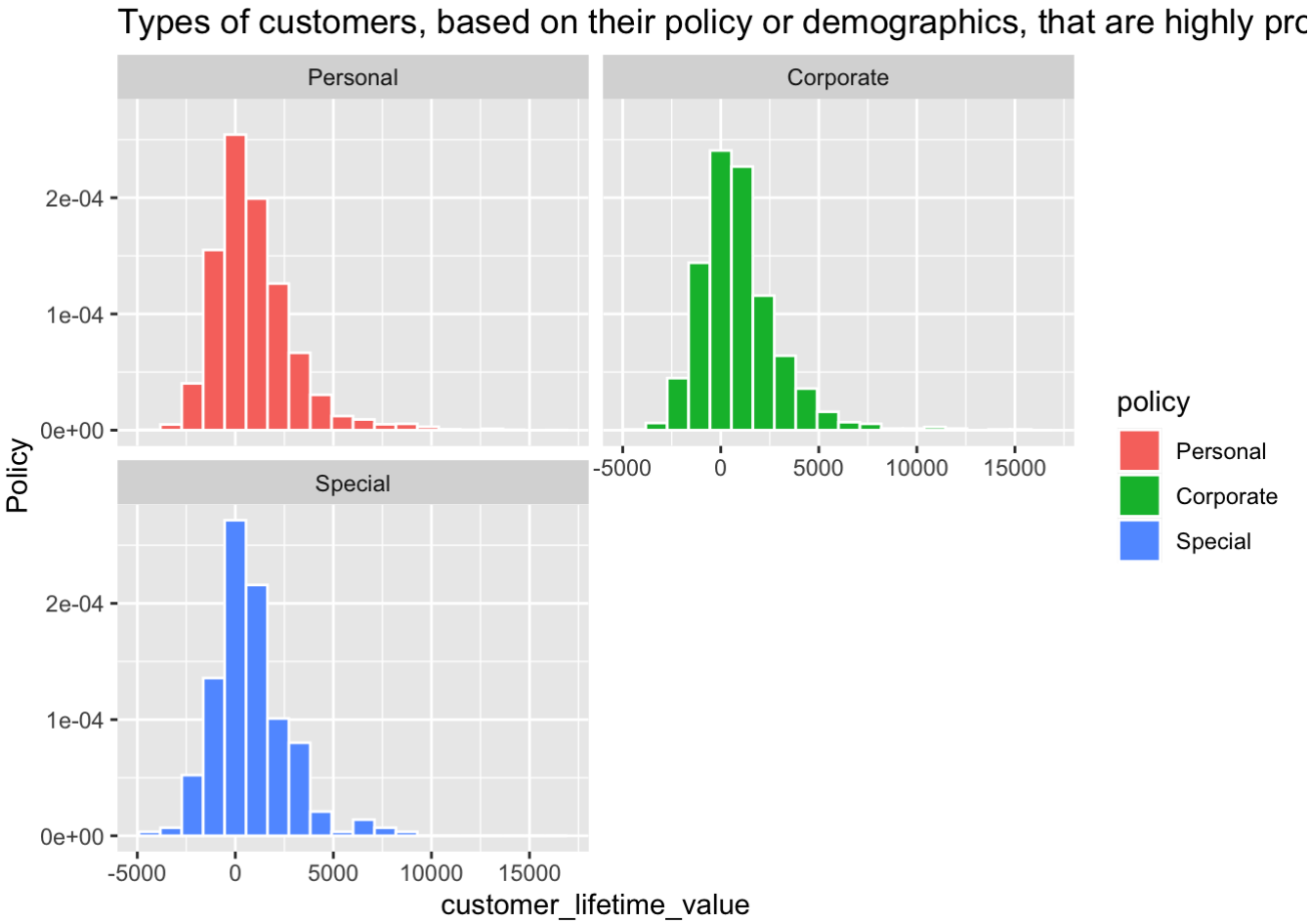
```
claims_df %>% group_by(policy) %>%
                  summarise(customer_id = n(),
                  min_customer_lifetime_value = min(customer_lifetime_value),
                  avg_customer_lifetime_value = mean(customer_lifetime_value),
                  max_customer_lifetime_value = max(customer_lifetime_value),
                  sd_customer_lifetime_value = sd(customer_lifetime_value))
```

policy <fct>	customer_id <int>	min_customer_lifetime_value <dbl>	avg_customer_lifetime_value <dbl>
Personal	4658	-4285	923.6647
Corporate	1328	-3850	951.3238
Special	263	-3911	745.9582

3 rows | 1-4 of 6 columns

#Data Visualization

```
ggplot(data = claims_df, aes(x = customer_lifetime_value, fill = policy)) +
   geom_histogram(aes(y = ..density..), color = "white", bins = 20) +
   facet_wrap(~ policy, nrow = 2) +
   labs(title = "Types of customers, based on their policy or demographics, that are hig
hly profitable",
           x = "customer_lifetime_value", y = "Policy")
```



Question: Do certain policies have a lower number of claims, leading to large profits?

Answer: Based on the findings, we can conclude that the average total number of claims for all policies is 2 and that there is a positive value for the average customer lifetime value. We also have greater profit when the total number of claims is 2 as opposed to 1,3, or 4. We can also observe that there is a significant loss when the total claims are four, as seen by the negative numbers for average customer lifetime value.

```
claims_df %>% group_by(policy,total_claims) %>%
                  summarise(customer_id = n(),
                  min_customer_lifetime_value = min(customer_lifetime_value),
                  avg_customer_lifetime_value = mean(customer_lifetime_value),
                  max_customer_lifetime_value = max(customer_lifetime_value),
                  sd_customer_lifetime_value = sd(customer_lifetime_value))
```

```
## `summarise()` has grouped output by 'policy'. You can override using the
## `.groups` argument.
```

policy	total_claims	customer_id	min_customer_lifetime_value	avg_customer_lifet
<fct>	<dbl>	<int>	<dbl>	
Personal	1	162	-1432	
Personal	2	2773	-2390	
Personal	3	1469	-3283	

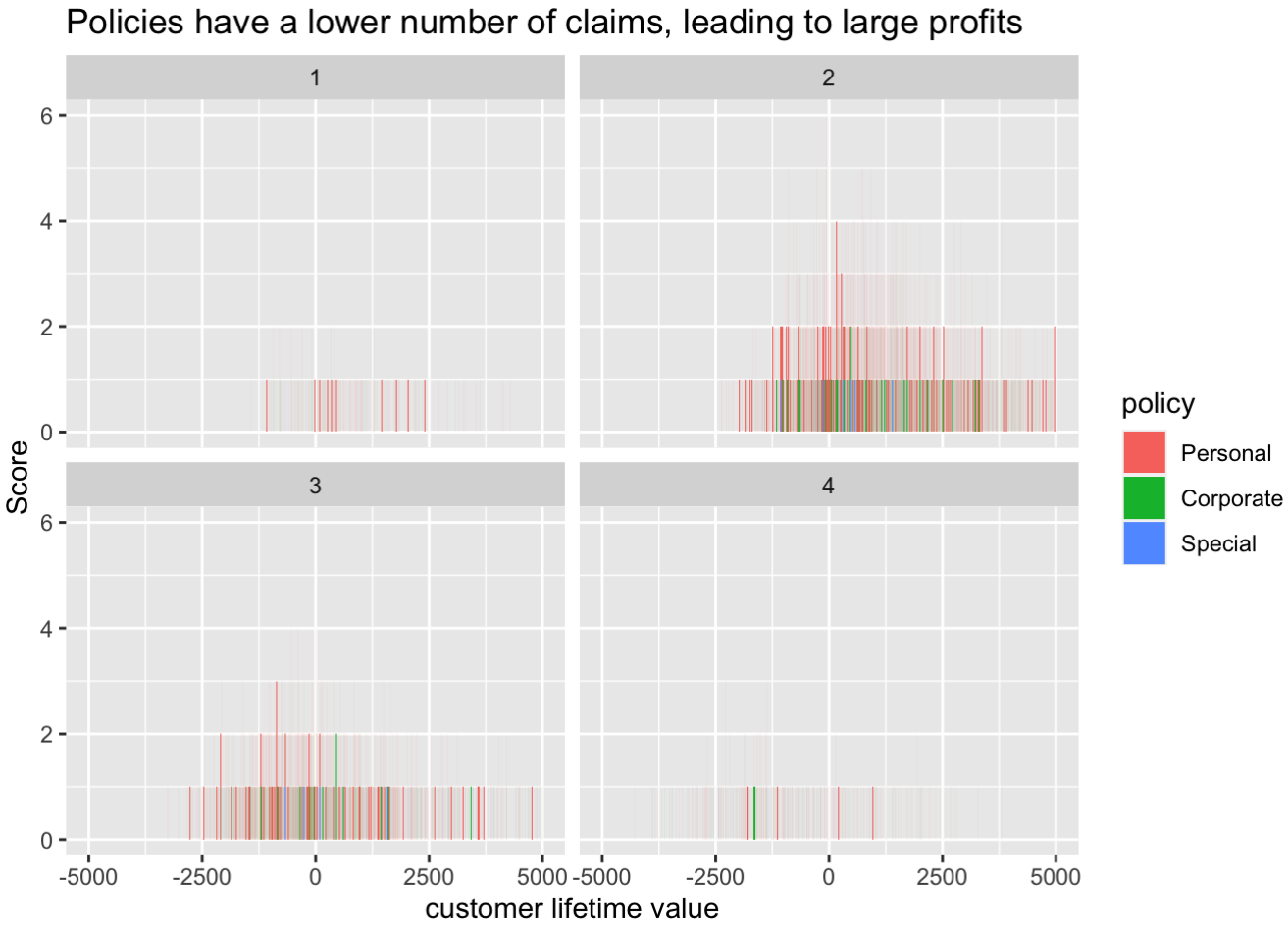
policy <fct>	total_claims <dbl>	customer_id <int>	min_customer_lifetime_value <dbl>	avg_customer_lifet
Personal	4	254	-4285	
Corporate	1	48	-941	
Corporate	2	786	-1844	
Corporate	3	421	-3029	
Corporate	4	73	-3850	
Special	1	9	-640	
Special	2	150	-1479	

1-10 of 12 rows | 1-5 of 7 columns Previous **1** 2 Next

#Data Visualization

```
ggplot(claims_df, aes(x=customer_lifetime_value,fill=policy))+
geom_bar(stat="count")+
  facet_wrap(~total_claims,nrow=2)+ xlim(-5000, 5000) +
 labs(x="customer lifetime value",
         y="Score",
       title="Policies have a lower number of claims, leading to large profits")
```

```
## Warning: Removed 242 rows containing non-finite values (stat_count).
```



Question: Does state effect the customer_lifetime_value ?

Answer: Yes, there is a link between customer states and customer lifetime value, as evidenced by the fact that when average total claims are high, average customer lifetime value is low, and vice versa, when average total claims are low, average customer lifetime value is high. State Nevada has high average customer lifetime value and Washington has least average customer lifetime value among the 5 states.

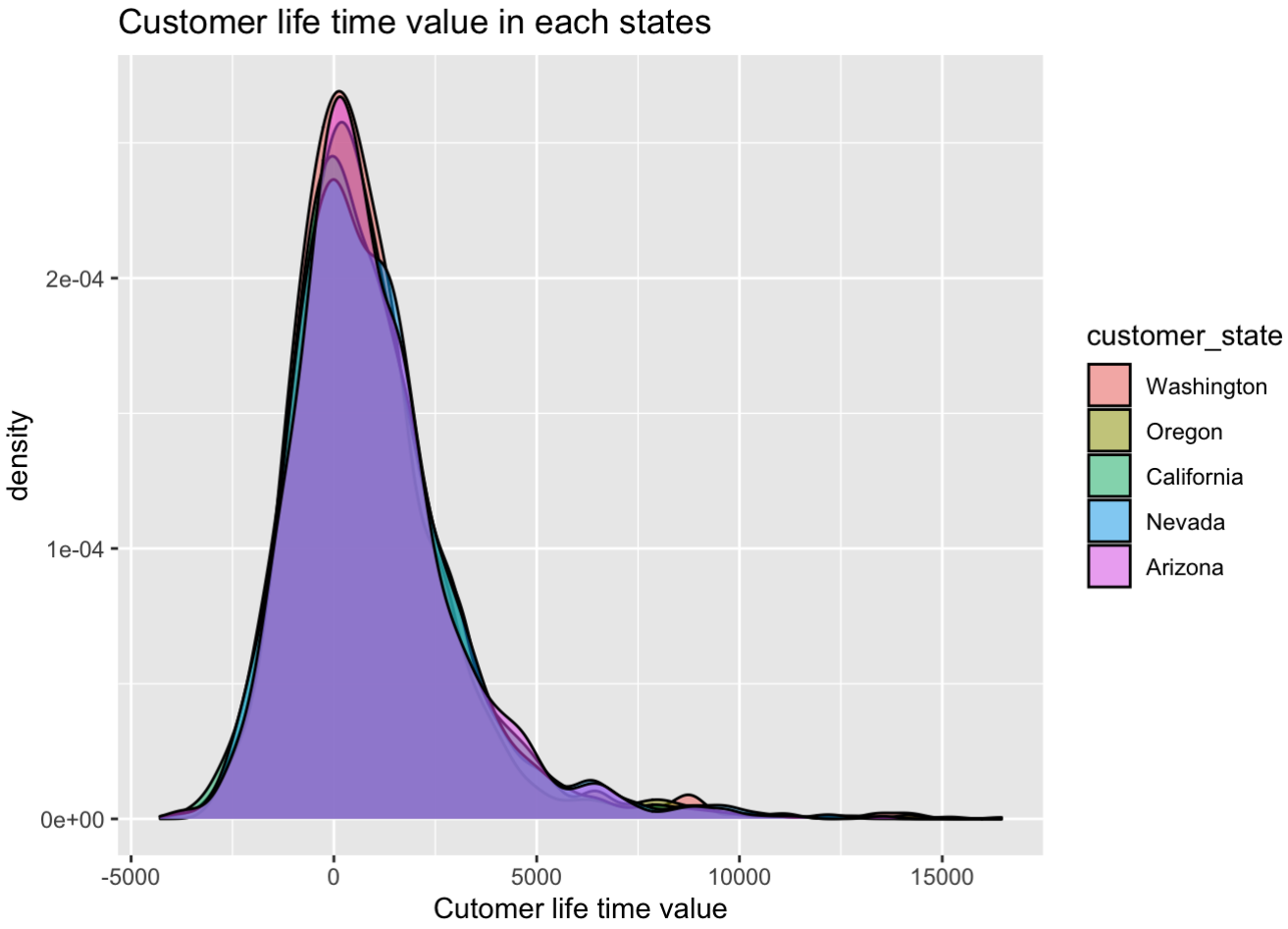
```
claims_df %>% group_by(customer_state) %>%
                  summarise(n_customers = n(),
                            avg_total_claims=mean(total_claims),
                            min_customer_lifetime_value = min(customer_lifetime_value),
                            avg_customer_lifetime_value = mean(customer_lifetime_value),
                            max_customer_lifetime_value = max(customer_lifetime_value),
                            sd_customer_lifetime_value = sd(customer_lifetime_value),)
```

customer_state <fct>	n_customers <int>	avg_total_claims <dbl>	min_customer_lifetime_value <dbl>
Washington	554	2.402527	-2741
Oregon	1763	2.384005	-4285
California	2150	2.411628	-3890
Nevada	601	2.384359	-3850
Arizona	1181	2.365792	-3911

5 rows | 1-4 of 7 columns

#Data Visualization

```
ggplot(claims_df, aes(x = customer_lifetime_value, fill = customer_state)) +
geom_density(alpha = 0.5) + ggtitle("Customer life time value in each states")+
  xlab("Cutomer life time value")+
  labs(fill = "customer_state")
```



Question: Is there a relationship between the vehicle_class and customer_lifetime_value?

Answer: Yes, there is a relationship.The average total claims for luxury SUVs are lower, and the average customer lifetime value is higher. The average total claim for a two-door automobile is high, while the average customer lifetime value is low. As a result, we may conclude that insurance companies benefit more from luxury automobiles and SUVs than from other types of vehicles.

```
claims_df %>% group_by(vehicle_class) %>%
                  summarise(n_customers = n(),
                            avg_total_claims=mean(total_claims),
                            min_customer_lifetime_value = min(customer_lifetime_value),
                            avg_customer_lifetime_value = mean(customer_lifetime_value),
                            max_customer_lifetime_value = max(customer_lifetime_value),
                            sd_customer_lifetime_value = sd(customer_lifetime_value))
```

vehicle_class <fct>	n_customers <int>	avg_total_claims <dbl>	min_customer_lifetime_value <dbl>
Two-Door Car	1292	2.411765	-3890
Four-Door Car	3124	2.387644	-4285
Sports Car	335	2.400000	-1740
SUV	1246	2.400482	-2934
Luxury Car	119	2.344538	60
Luxury SUV	133	2.233083	-287

6 rows | 1-4 of 7 columns

Question: Is there a link between customer lifetime value and employee status?

Answer: Despite the fact that disabled employment status has a higher average of total claims, disability employment status has a higher average customer lifetime value. Customers that have employed as their employment status have a high Standard Deviation.

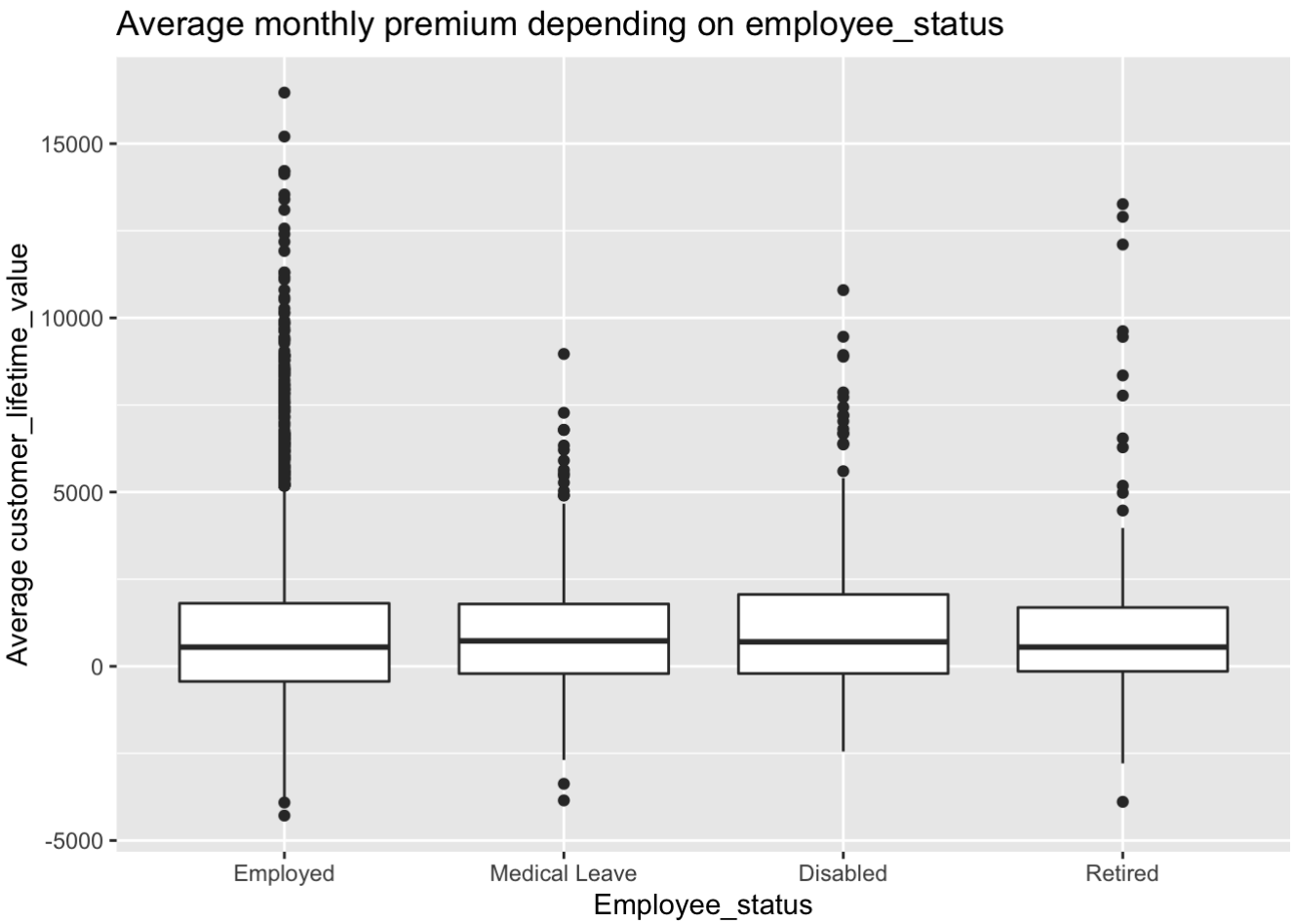
```
claims_df %>% group_by(employment_status) %>%
                  summarise(n_customers = n(),
                            avg_total_claims=mean(total_claims),
                            min_customer_lifetime_value = min(customer_lifetime_value),
                            avg_customer_lifetime_value= mean(customer_lifetime_value),
                            max_customer_lifetime_value = max(customer_lifetime_value),
                            sd_customer_lifetime_value = sd(customer_lifetime_value))
```

employment_status <fct>	n_customers <int>	avg_total_claims <dbl>	min_customer_lifetime_value <dbl>
Employed	5154	2.387466	-4285
Medical Leave	421	2.418052	-3850
Disabled	392	2.443878	-2447
Retired	282	2.358156	-3890

4 rows | 1-4 of 7 columns

#Data Visualization

```
ggplot(claims_df, aes(x=customer_lifetime_value,y = employment_status
,fill=customer_lifetime_value)) + geom_boxplot() + coord_flip()+
  labs(title = "Average monthly premium depending on employee_status ",
           x = "Average customer_lifetime_value",
           y = "Employee_status")
```



Question: Does sales_channel at the company related to him/her customer_lifetime_value?

Answer: Clients who came to the insurance business through an agent had a lower average total claim and a higher average customer lifetime value. The majority of Aslo's customers came through agents.

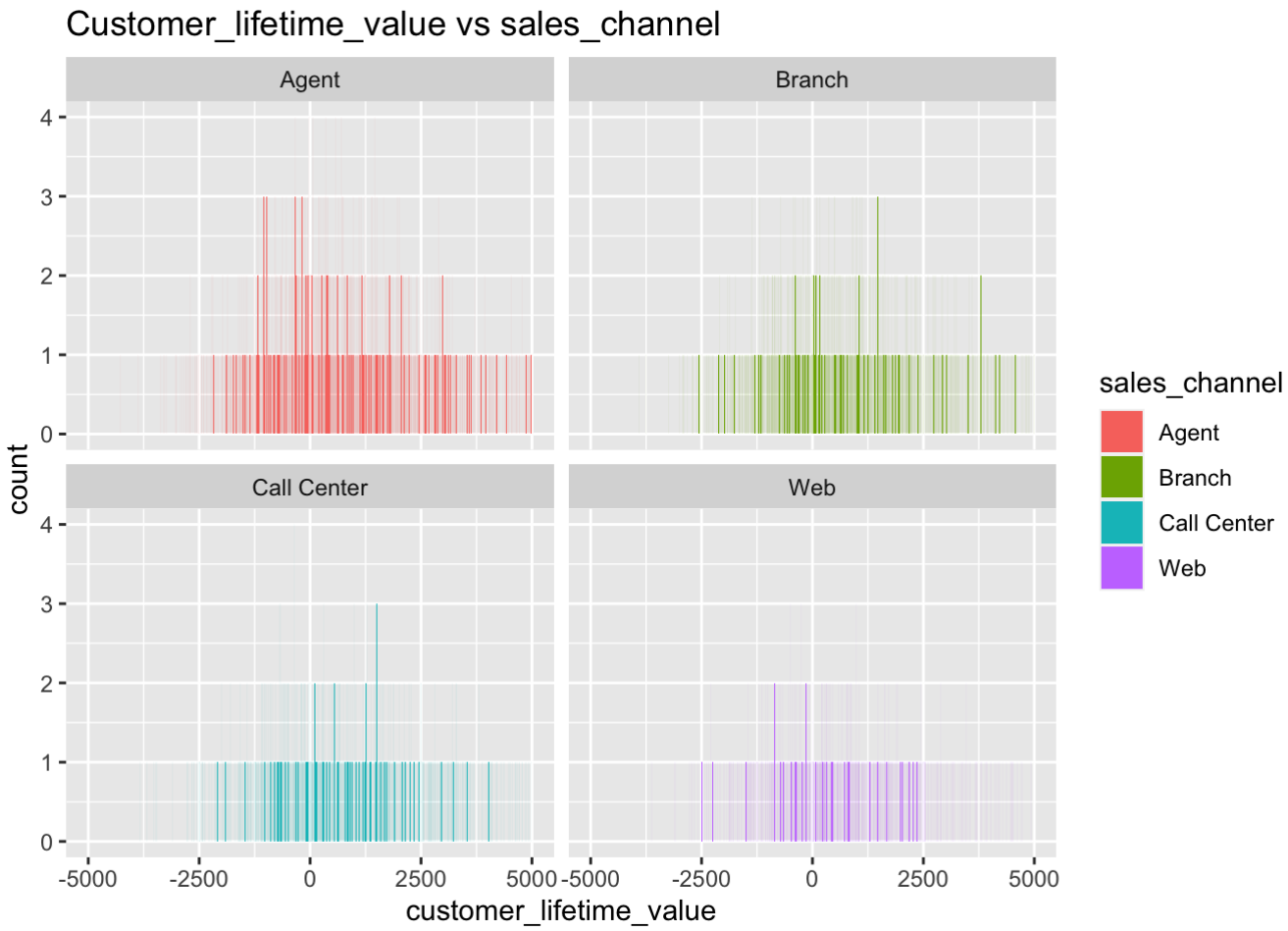
```
claims_df %>% group_by(sales_channel) %>%
                  summarise(n_customers = n(),
                            avg_total_claims=mean(total_claims),
                            avg_customer_lifetime_value= mean(customer_lifetime_value))
```

sales_channel	n_customers	avg_total_claims	avg_customer_lifetime_value
<fct>	<int>	<dbl>	<dbl>
Agent	2359	2.380670	969.9640
Branch	1771	2.395257	888.0265
Call Center	1218	2.393268	895.6658
Web	901	2.411765	899.2386

4 rows

```
ggplot(data = claims_df, aes(x = customer_lifetime_value, fill = sales_channel)) +
       geom_bar() + facet_wrap(~ sales_channel, nrow = 2) +xlim(-5000,5000) +
       labs(title = "Customer_lifetime_value vs sales_channel",
            x = "customer_lifetime_value",
            y = "count")
```

```
## Warning: Removed 242 rows containing non-finite values (stat_count).
```



Question: How residence_type are related to customer_lifetime_value?

Answer: According to the statistics, the average monthly policy active for rural residences is high, while the average customer lifetime value for suburban residences is high.

```
claims_df %>% group_by(residence_type) %>%
                  summarise(n_customers = n(),
                            min_months_policy_active = min(months_policy_active),
                            avg_months_policy_active = mean(months_policy_active),
                            max_months_policy_active = max(months_policy_active),
                            sd_months_policy_active = sd(months_policy_active))
```

residence_type	n_customers	min_months_policy_active	avg_months_policy_active
<fct>	<int>	<dbl>	<dbl>
Urban	1495	12	38.73177

residence_type <fct>	n_customers <int>	min_months_policy_active <dbl>	avg_months_policy_active <dbl>
Suburban	3657	12	39.07055
Rural	1097	12	39.29353

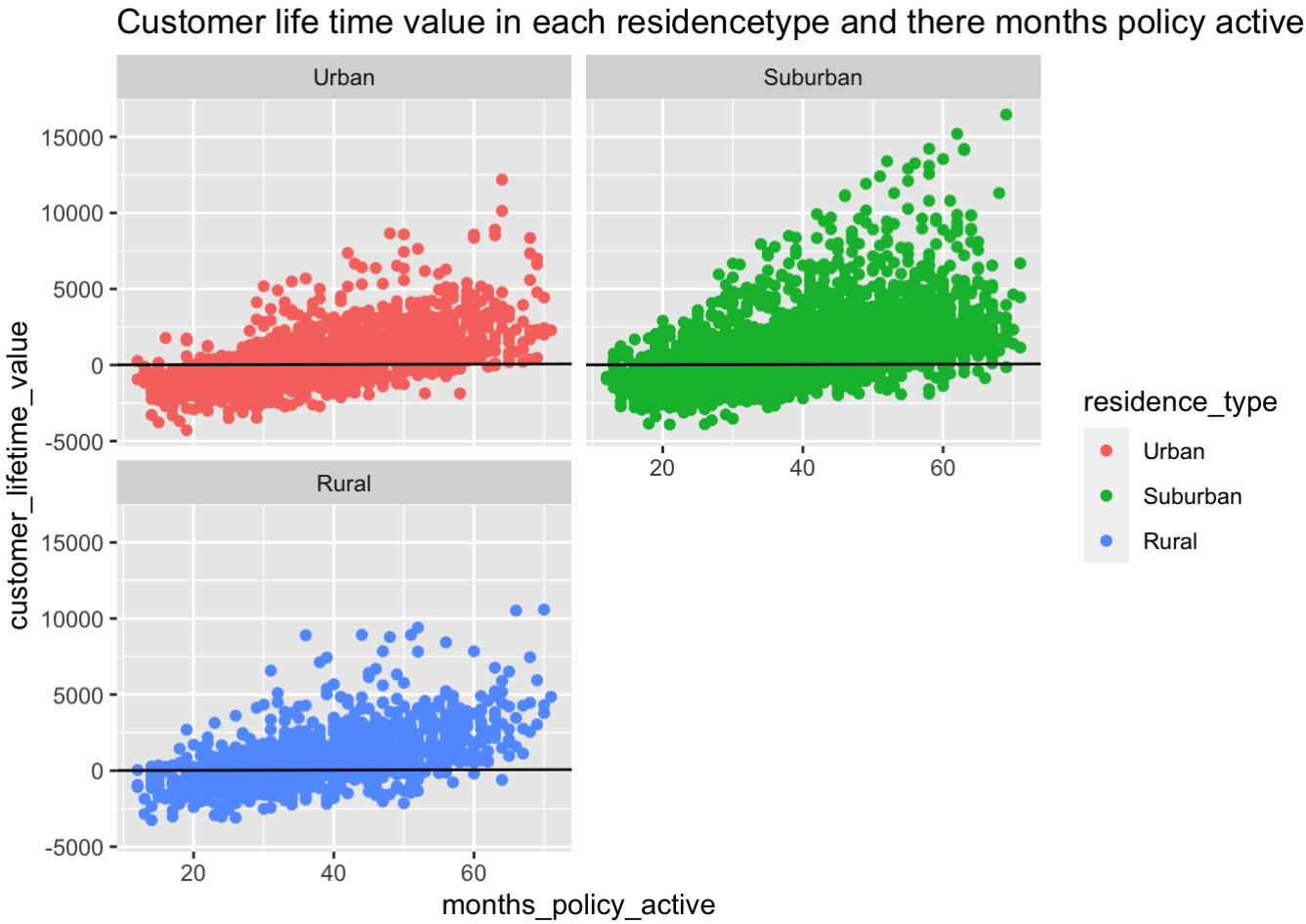
3 rows | 1-4 of 6 columns

```
claims_df %>% group_by(residence_type) %>%
                  summarise(n_customers = n(),
                            min_customer_lifetime_value = min(customer_lifetime_value),
                            avg_customer_lifetime_value= mean(customer_lifetime_value),
                            max_customer_lifetime_value = max(customer_lifetime_value),
                           sd_customer_lifetime_value = sd(customer_lifetime_value))
```

residence_type <fct>	n_customers <int>	min_customer_lifetime_value <dbl>	avg_customer_lifetime_value <dbl>
Urban	1495	-4285	563.0488
Suburban	3657	-3911	1055.5710
Rural	1097	-3262	966.2662

3 rows | 1-4 of 6 columns

```
ggplot(claims_df, aes(x= months_policy_active , y= customer_lifetime_value)) +geom_point
(aes(colour= residence_type))+ facet_wrap(~residence_type,nrow=2)+geom_abline()+ggtitle(
"Customer life time value in each residencetype and there months policy active")
```



Question: Is there a relationship between marital status and customer_lifetime_value?

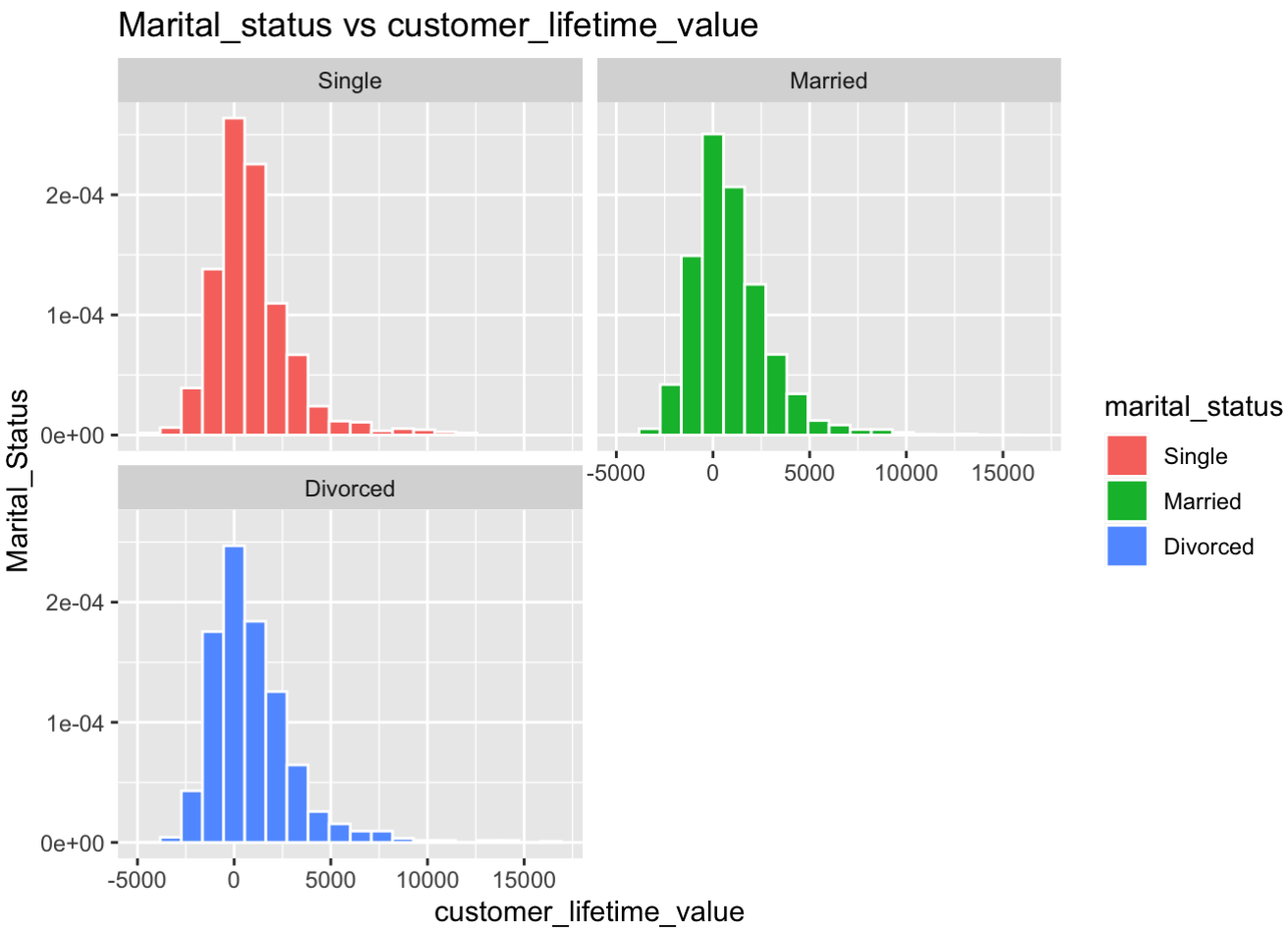
Answer: Yes, there is a link between the two. Customers with the marital status of single have a greater average customer life value than customers with the marital status of Married and Divorced. Customers with the marital status of married have a lower average customer life time value.

```
claims_df %>% group_by(marital_status) %>%
                  summarise(n_customers = n(),
                            min_customer_lifetime_value = min(customer_lifetime_value),
                            avg_customer_lifetime_value= mean(customer_lifetime_value),
                            max_customer_lifetime_value = max(customer_lifetime_value),
                            sd_customer_lifetime_value = sd(customer_lifetime_value))
```

marital_status	n_customers	min_customer_lifetime_value	avg_customer_lifetime_value
<fct>	<int>	<dbl>	<dbl>
Single	1027	-4285	947.1052
Married	4158	-3911	915.0200
Divorced	1064	-3112	925.4182

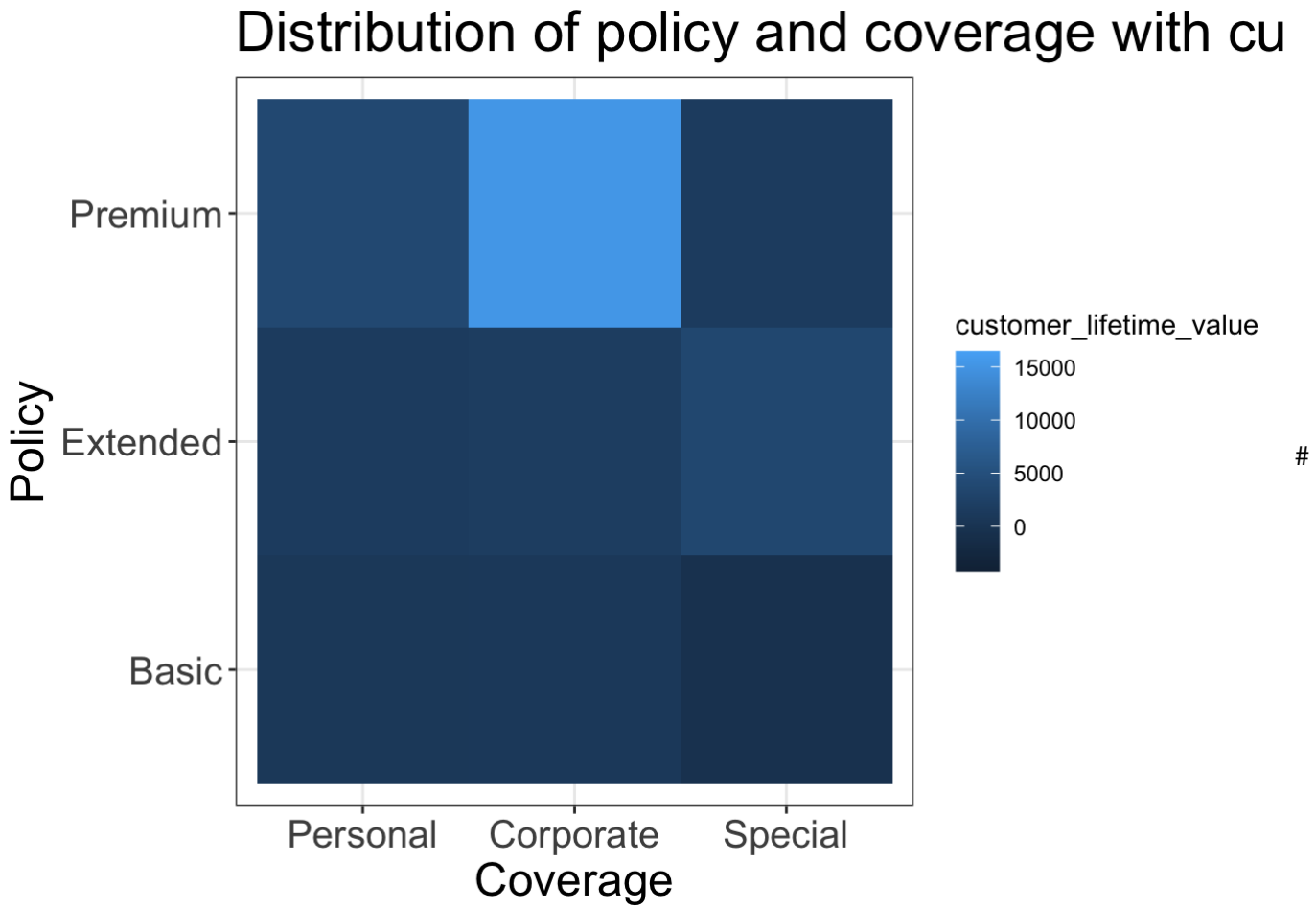
3 rows | 1-4 of 6 columns

```
ggplot(data = claims_df, aes(x = customer_lifetime_value, fill = marital_status)) +
   geom_histogram(aes(y = ..density..), color = "white", bins = 20) +
   facet_wrap(~ marital_status, nrow = 2) +
   labs(title = "Marital_status vs customer_lifetime_value",
           x = "customer_lifetime_value", y = "Marital_Status")
```



#Data Visualization of distribution of policy and coverage with customer_lifetime_value.

```
ggplot(sample_n(claims_df,6249), aes(policy,coverage, fill=customer_lifetime_value)) +
  geom_tile()+
labs(y="Policy",
        x="Coverage",
      title="Distribution of policy and coverage with customer_lifetime_value)")+
theme_bw()+
theme(plot.title = element_text(size=22)
      ,axis.text.x= element_text(size=15),
       axis.text.y= element_text(size=15),
        axis.title=element_text(size=18))
```



Summary of Results

Write an executive summary of your overall findings and recommendations to the executives at this company. Think of this section as your closing remarks of a presentation, where you summarize your key findings and make recommendations to improve pricing, company operations, and car insurance policy adjustments.

Your executive summary must be written in a professional tone (https://www.universalclass.com/articles/writing/business-writing/appropriate-tone-in-business-communications.htm), with minimal grammatical errors, and should include the following sections:

1. An introduction where you explain the business problem and goals of your data analysis
    - What problem(s) is this company trying to solve? Why are they important to their future success?
    - What was the goal of your analysis? What questions were you trying to answer and why do they matter?

2. Highlights and key findings from your Exploratory Data Analysis section
    - What were the interesting findings from your analysis and **why are they important for the business**?
    - This section is meant to **establish the need for your recommendations** in the following section

3. Your recommendations to the company
    - Each recommendation must be supported by your data analysis results

- You must clearly explain **why** you are making each recommendation and which results from your data analysis support this recommendation
- You must also describe the potential business impact of your recommendation:
  - Why is this a good recommendation?
  - What benefits will the business achieve?

Executive Summary

Please write your executive summary below. If you prefer, you can type your summary in a text editor, such as Microsoft Word, and paste your final text here.

1.Customers play a critical role in their individual businesses. So the key goal is to figure out how the firm must balance total income and total claims in order to make a profit. It's good to know that we're extracting information from the data to make it simpler to make decisions about how to grow the company's revenues.So it is important to understand and analyze all the factors so that it can bring a change in the company and insurance attrition. The questions that were answered in my analysis are:

Are there types of customers, based on their policy or demographics, that are highly profitable? Do certain policies have a lower number of claims, leading to large profits? *Does state effect the customer_lifetime_value ?* Is there a relationship between the vehicle_class and customer_lifetime_value? *Is there a link between customer lifetime value and employee status?* Does sales_channel at the company related to him/her customer_lifetime_value? *How residence_type are related to customer_lifetime_value?* Is there a relationship between marital status and customer_lifetime_value?

2. There are some crucial findings that have been gathered from the analysis. They are: The intriguing thing we see is that the majority of our clients opt for personal insurance. However,the average customer life value is larger than the other two plans, insurance firms profit from corporate policyholders. The average total number of claims across all plans is two, and the average customer lifetime value is positive. We also make more money when the overall number of claims is two rather than one, three, or four. We can also see that when the total claims are four, there is a considerable loss, as seen by the negative statistics for average customer lifetime value. The fact that when average total claims are high, average customer lifetime value is low, and vice versa, when average total claims are low, average customer lifetime value is high, shows that there is a relationship bet ween customer state and customer lifetime value.State Nevada has high average customer lifetime value and Washington has least average customer lifetime value among the 5 states. Luxury SUVs have a lower average total claim value and a higher average customer lifetime value. For a two-door car, the average total claim is high, while the average customer lifetime value is low. As a consequence, we may deduce that expensive cars and SUVs benefit insurance firms more than other types of vehicles. Disability employment status has a greater average customer lifetime value than non-disabled work status, while having a larger average of total claims. The Standard Deviation of customers who have employed as their employment status is high. Rural residents have a high average monthly policy active rate, whereas suburban residents have a high average customer lifetime value. Customers who are single have a higher average customer life value than customers who are married and divorced.Customers who are married have a lower average customer lifetime value than those who are single. The average total claim was lower and the average customer lifetime value was higher for clients who arrived to the insurance firm through an agent. Customers were mostly acquired through agents.

3.Even though we have most of the customers from personal policy but we get profits from corporate policy. As insurance providers we may have to increase the monthly premium for the personal policy. As the number of claims increases there should be increase in the certain percentage of monthly premium, so as to balance the

total revenue and total claims amount. Increase in monthly premium in the states where the average total claims might increase the profit or balance the total revenue and total claim amount. We know that states like Washington and California has highest accident rate therefore max claim amount. Luxury Cars have more monthly premium but the average customer lifetime values is more for them and we are getting profits from them. May be can try decreasing the monthly premium to a minor percentage and attract more number of customers by showing the sale or discount offer. We can increase the monthly premium depending on the resident_type(i.e, Rural,Urban, and Suburban) such as in urban areas and suburban areas we can increase the monthly_premium compared with rural area type.